Pacific Symposium on Biocomputing 15:444-455(2010)

CLUSTERING CONTEXT-SPECIFIC GENE REGULATORY NETWORKS*

ARCHANA RAMESH, ROBERT TREVINO

School of Computing, Informatics & Decision Systems Engineering,
Arizona State University, 699 S Mill Avenue, Tempe, AZ 85281, USA

DANIEL D. VON HOFF

Clinical Translational Research Division, Translational Genomics Research Institute,
445 North Fifth Street, Phoenix, AZ 85004, USA

SEUNGCHAN KIM

School of Computing, Informatics & Decision Systems Engineering,
Arizona State University, 699 S Mill Avenue, Tempe, AZ 85281, USA
Computational Biology Division, Translational Genomics Research Institute,
445 North Fifth Street, Phoenix, AZ 85004, USA
E-mail: dolchan@asu.edu

Gene regulatory networks (GRNs) learned from high throughput genomic data are often hard to visualize due to the large number of nodes and edges involved, rendering them difficult to appreciate. This becomes an important issue when modular structures are inherent in the inferred networks, such as in the recently proposed context-specific GRNs.[12] In this study, we investigate the application of graph clustering techniques to discern modularity in such highly complex graphs, focusing on context-specific GRNs. Identified modules are then associated with a subset of samples and the key pathways enriched in the module. Specifically, we study the use of Markov clustering and spectral clustering on cancer datasets to yield evidence on the possible association amongst different tumor types. Two sets of gene expression profiling data were analyzed to reveal context-specificity as well as modularity in genomic regulations.

Keywords: Markov clustering; spectral clustering; cancer; gene regulatory networks; cellular context

1. Introduction

A cell maintains a specific state (such as 'healthy') by tightly regulating a set of molecules. When exposed to environmental changes, the cell adjusts its regulatory mechanisms and transitions to a state (such as 'tumor') significantly different from the original state. Since the manner in which the system reacts to inputs is altered, we term this as a change in cellular context.[8]

Kim et al.[12] have proposed an algorithm which uses a probabilistic framework to learn contexts from gene expression data. More recently, Sen et al.[19] have applied this method to identify context-specific gene regulatory networks (GRNs). Unlike conventional GRNs, edges in context-specific GRNs represent the interaction conditioned on a subset of samples, i.e. *their biological context*, thus lending adaptability to the model of biological regulation.

However, GRNs learned by the algorithm are often made of a few thousand nodes (genes) and tens of thousands of interactions rendering manual curation of the edges and sub-network to identify its modular structure and context-specificity quite difficult if not impossible. Hence, the need for the automatic extraction and curation of relevant *context clusters* from these networks is critical.

Graph clustering is defined as the task of grouping the vertices of a graph in such a way that there are many edges within a cluster and relatively few edges between the clusters.[18] The most significant difference between conventional clustering and graph clustering is in the notion of the relationship between the elements being clustered. When similarity is expressed through whether elements "share a property" or not (such as a regulatory relationship where genes are co-regulated), rather than the distance between the elements, graph

[*] Code, scripts and supplementary information available online at http://sysbio.fulton.asu.edu/publications/2010/psb2010/

clustering is appropriate for the problem. Moreover, as the problem of clustering GRNs is also directly related to the connectivity between nodes, we believe that graph clustering is better suited for solving this problem.

Our work looks into the applicability of two graph clustering algorithms, namely Markov clustering and spectral clustering in identifying the modular structure of context-specific GRNs. Markov clustering was chosen due to its scalability and ability to automatically determine the number of clusters. Spectral clustering was chosen due to its ability to find an optimal minimum cut while creating well-balanced clusters. In addition, previous applications in the bioinformatics field have yielded promising results,[7,10,17,23] leading us to believe it would be well-suited for this problem.

Our paper is organized as follows. We begin with an overview of the existing applications of graph clustering to bioinformatics. Following this, we provide a mathematical formulation of our problem and then describe the graph clustering methods and enrichment analysis techniques that we apply. Subsequently we demonstrate how our methods could be applied to yield insights on the underlying mechanisms of cancer. Finally, we conclude with the future direction of our work.

2. Relevant Work

Clustering is defined as the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters).[11] The clustering task usually involves pattern representation, definition of pattern proximity, clustering or grouping, data abstraction and assessment.[5]

Both Markov clustering and spectral clustering have been previously applied to bioinformatics. Lattimore et al.[17] have applied MCL to the analysis of microarray data using a graph constructed from the correlation of gene expression measurement to which MCL is applied. The algorithm has been applied to a breast cancer dataset and shown to identify underlying biological mechanisms. In a similar study,[7] Freeman et al. have used MCL in the clustering and visualization of transcription networks from microarray data. This work focuses on applying MCL to transcription networks derived from mouse gene expression data, again using the Pearson's correlation coefficient.

Spectral clustering has been used in the analysis of gene expression data. Clustering has been performed in terms of genes, samples and both dimensions. While Tritchler et al.[23] apply Eigen analysis (spectral methods) to cluster genes from gene expression data, Higham et al.[10] apply spectral methods to cluster gene expression data based on samples. Kluger et al.[13] have shown that the eigenvectors in matrices of gene expression data formed a distinct "checkerboard" pattern that can be exploited to simultaneously cluster genes and conditions in cancer datasets. In all cases, while our method shares similar objectives, our work differs in the sense that we apply spectral clustering to the extracted context-specific gene regulatory networks as a graph clustering approach.

The primary contributions of our work include developing methods for clustering the recently proposed context-specific gene regulatory networks. Context-specific gene regulatory networks provide a means to specify genomic regulations conditioned by a subset of samples. Secondly, our work focuses on comparisons between Markov clustering and two variants of Spectral clustering, suited for both symmetric and asymmetric graphs. Comparisons are performed in two dimensions- using performance measures such as coverage and performance which measure the goodness of the obtained clusters and using enrichment analysis which allows for the biological interpretation of the results. Finally, both algorithms are applied to two cancer gene expression datasets yielding insights on possible associations between tumor-types and several useful clinical implications.

3. Problem Definition

A cellular context is defined as a set of genes, one or more of which function as drivers and the rest as driven genes, exhibiting consistent transcriptional behavior across a set of samples, drawn from a cellular process governed by tightly regulated mechanism(s) involving the set of genes. Mathematically, a context can be represented as $C_i = (G_i, Y_i, S_i, M_i)$ where G_i represents a set of driver genes, Y_i represents the possible states of the genes (an example would be $\{-1, 0, +1\}$ for a ternary quantized dataset), S_i represents a set of driven genes and M_i represents the set of samples under which consistent expression is observed.

Each context defines regulatory relationships between the driver genes and the driven genes, i.e. $G_i \rightarrow g \in S_i$, specific to M_i with G_i (drivers) conditioned on a specific state $Y_i = y_i$. A driver g_j in context C_j could be driven by g_i in another context C_i. When such relationships are added to the implicit driver-driven relationships $g_i \rightarrow g_j$, we obtain an interesting graphical structure representing the relationships between contexts. We call this graph, a *context-specific GRN* as each regulatory relationship $g_i \rightarrow g \in S_i$ is specific to a subset of samples M_i.

In graph theoretic terms, a context-specific GRN is a directed graph $G = (V, E)$ where V is a set of vertices representing genes and E is a set of edges representing context-specific driver-driven relationships.

A partition of the graph into two non-empty sets S and $V \setminus S$ is called a *cut* and denoted by $(S, V \setminus S)$. Usually a cut is uniquely defined by a set S, and hence any sub-set of V can be called a cut. The *cut-size* is the number of edges that connects vertices in S to those in $V \setminus S$.

Given a context-specific GRN $G = (V, E)$ as defined above, our goal is to determine the clusters of the network; where a cluster C may be defined as an induced subgraph of the graph, such that $C = (V_c, E_c)$, where $V_c \in V, E_c \in E$; (i) for every edge $(u, v) \in E_c$, $u \in V_c$ and $v \in V_c$ and (ii) the cut size of the cluster C is minimal.[18]

4. Methods

Contexts are learned from gene expression data through the cellular context mining algorithm.[12] Given a gene g_k and a cellular context c_j defined by a subset of samples M_j, the algorithm uses probabilistic measures to identify a set of genes with consistent expression levels within the context. The resulting contexts dictate implicit driver-driven relationships. When these relationships are captured in the form of a graph, we obtain a context-specific GRN.[19] In our study, we used a variant of this context-specific GRN, where each node was the driver of a context. The set of genes regulated by driver S_j was used in the enrichment analysis.

4.1. *Markov Clustering*

Markov clustering derives its inspiration from the notion of random walks in graphs. If a random walk visits a node in a cluster, it would be likely to visit several other members of the cluster before leaving the cluster.[24]

The Markov clustering algorithm simulates flow using two (alternating) algebraic operations on matrices. Expansion (identical to matrix multiplication) represents the homogenization of flow across different regions of the graph. Inflation, mathematically equivalent to a Hadamard power followed by diagonal scaling, represents the contraction of flow, making it thicker in regions of higher current and thinner in regions of lower current. Intuitively, expansion corresponds to augmenting the neighbors of a given vertex, and inflation corresponds to promoting those neighbors which have a higher transition probability from a given vertex. The MCL process causes flow to spread out within natural clusters and disappear in between different clusters.[24] The iteration is continued until a recurrent state or fixpoint is reached. The exact steps are explained in Algorithm 4.1. The connected components of the graph induced by the non-zero entries of M provide the required clustering. Proof of concept, mathematical properties and analyses on the complexity and scalability of the algorithm can be found in van Dongen's work.[25] Our implementation of Markov clustering used the publicly available tool BioLayout Express.[7]

4.2. *Spectral Clustering*

Spectral clustering uses the Eigen decomposition of matrix representations of a graph to determine the optimal partitioning of the graph. Although, there has been extensive research in the spectral clustering field, we used the algorithms developed by Shi and Malik[20] and Meila and Pentney[15] because they incorporate information from the edges (in our case, computationally predicted biological interactions) in determining the optimal clustering of a graph.

Algorithm 4.1 Markov Clustering

```
Input: $G = (V, E)$, expansion parameter $e$, inflation parameter $r$
while $M$ is not fixpoint do
  $M \leftarrow M^e$
  for all $u \in V$ do
    for all $v \in V$ do
      $M_{uv} \leftarrow M_{uv}^r$
    end for
    for all $v \in V$ do
      $M_{uv} \leftarrow \frac{M_{uv}}{\sum_{w \in V} M_{uw}}$
    end for
  end for
end while
```

4.2.1. *Symmetric Cuts:*

In graph theory, a cut is defined as

$$cut(A, B) = \sum_{u \in A, v \in B} w_{uv}, \tag{1}$$

where A and B are the clusters resulting from the cut between vertices u and v. Finding the minimum cut for Equation 1 could result in singletons or clusters with very few nodes, leading to poorly distributed clusters. Thus, there exists a need to balance the clusters. Shi and Malik, have proposed a solution to this problem by normalizing the cuts that create clusters.[20] The cut cost is calculated as a fraction of the weights of the edges in the induced sub-graphs. As finding the exact solution to the normalized minimum cut problem is considered NP-complete, the authors have found that using the eigenvector corresponding to the second smallest eigenvalue of the Laplacian of an undirected graph (also known as the Fiedler vector) could efficiently provide an approximate discrete solution.[20] The algorithm, referred to as the normalized cut algorithm, recursively splits clusters thresholding the Fiedler vector of the induced sub-graphs until the desired number of clusters are reached.

4.2.2. *Asymmetric Cuts:*

Meila and Pentney[15] provide for the expansion of spectral clustering in multi-way cuts to directed graphs, as the normalized cut is applicable only to undirected graphs. In gene regulation directionality could provide useful information. The weighted cut algorithm, proposed by Meila and Pentney, mathematically transforms a directed graph (with a non-normalized Laplacian matrix, D-A), into a symmetric Hermitian matrix[15] and finds an approximate solution to minimizing a normalized cut. Using the k eigenvectors pertaining to the k smallest eigenvalues of the Hermitian matrix, the weighted cut algorithm applies the k-means algorithm to cluster the graph. In addition, the algorithm allows for user input, balancing parameters T and T' , to normalize the cuts produced by the algorithm. Thus the normalized minimum cut for directed graphs can be expressed as:

$$MNCut(x) = \min_{z_k \in R^n \text{orthon}} \sum_{k=1}^{K} z_k^* H(B) z_k \tag{2}$$

where $B = T^{-\frac{1}{2}}(D - A)T^{-\frac{1}{2}}$, K is the number of desired clusters and $H(B)$ is the Hermitian matrix of B.

4.3. *Enrichment Analysis*

Subsequent to clustering the context-specific GRNs, it is interesting to study the pathways and phenotypic characteristics that the resulting clusters are enriched with. To this end, we employ the following mechanisms

Pacific Symposium on Biocomputing 15:444-455(2010)

to evaluate the biological significance of the obtained clusters.

4.3.1. *Gene Set Enrichment Analysis:*

We investigate the enrichment of each context cluster using gene sets. The hypergeometric test is used to measure the significance of the enrichment and the p-values are corrected for False Discovery Rate (FDR) using Benjamini and Hochberg's method.[1] The Molecular Signatures Database (MSigDB) is used as a reference knowledge source.[22] MSigDB contains a collection of gene sets, including positional gene sets, curated pathways, conserved motifs, computationally predicted expression neighborhoods (defined on 380 cancer-associated genes) and Gene Ontology gene sets.

4.3.2. *Tumor Type Enrichment Analysis:*

Sample Association Score: As the context clusters derived from the clustering process consist of a set of cellular contexts, it is relevant to study the samples that occur in more than one context within each context cluster. Samples were scored based on their occurrence within each context, over all the contexts within the cluster. A sample s, given a context cluster CC with m contexts $C_1, C_2, \cdots, C_m$, would have the scoring

$$\mathrm{SAS}(s, CC) = \sqrt[m]{\prod_{i=1}^{m} f_i(s)}, \tag{3}$$

where $f_i(s) = k_i/N$, when $s \in C_i$ and 1, otherwise.

Motivated by the differences in the number of samples in each context, a sample belonging to a larger context would have a lesser contribution to the score than a sample belonging to a smaller context. Only samples that had a sample association score < 0.5 were considered. Following this, the context clusters were analyzed for enrichment of specific tumor types using the Hypergeometric test. FDR correction was applied using Benjamini and Hochberg's correction method.[1]

5. Results

We applied our methods to two gene expression datasets – the Target Now dataset and the REMBRANDT study. In the following section, we discuss the study that was conducted, the results obtained including biological significance and performance comparisons of the three algorithms.

5.1. *Target Now Data*

Our input graph constituted a variant of the context-specific GRN produced by Sen et al.[19] from the Target Now (TN) dataset; a study aimed at determining if patients with refractory cancer, who did not benefit from the standard types of treatment, could derive benefit from therapy with a drug not normally used for their particular form of cancer.[26]

The dataset consists of 17,085 unique probes (Agilent-011521 Human 1A Microarray G4110A) from 146 patients with different types of refractory cancer. We used the graphs corresponding to the relationships derived from statistically significant contexts (using a p-value < 0.001). The graph consisted of 391 contexts and was organized into six strongly connected components.

5.1.1. *Markov Clustering*

As Markov clustering has a propensity towards undirected graphs, we used the undirected version of the context-specific GRN obtained from the filtered contexts. Clustering was performed on the graph using an inflation value of 2.0. The inflation parameter is used to control the granularity of the clusters obtained, and was set to 2.0 as it provided the desired granularity. Clusters with less than 3 nodes were not considered.

Pacific Symposium on Biocomputing 15:444-455(2010)

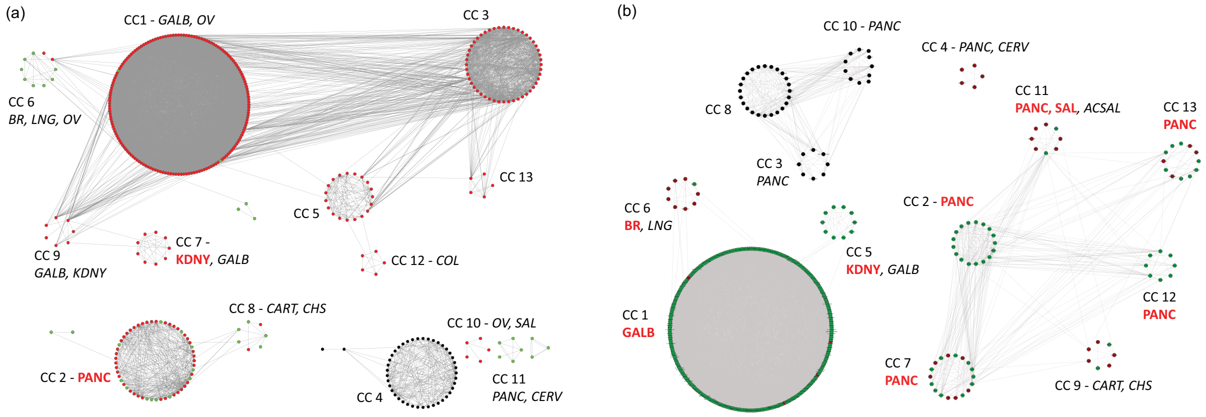

Fig. 1: (a) Markov Clustering Results [TN study]. (b) Asymmetric Spectral Clustering Results [TN study]. The following acronyms indicate enriched tumor types - BR:Breast, LNG:Lung, OV:Ovarian, PANC:Pancreas, GALB:Gall Bladder, CART:Cartilage, CHS:Chondrosarcoma, COL:Colon, SAL:Salivary, KDNY:Kidney. Tissue types in red indicates the type is over-represented in the corresponding context-cluster with *adjusted* p-value < 0.05. Tissue types with italicized font indicate the tissue type is over-represented in the corresponding context-cluster with p-value < 0.05.

The clusters obtained are shown in Figure 1(a). As seen in the illustration, the algorithm identified thirteen distinct clusters, dividing the six strongly connected components into smaller clusters. Further, we also note that within most clusters, the expression levels of all genes belonging to the cluster are similar.

Nine out of thirteen clusters were found to be enriched with several tumor types. Further discussion of the biological significance of these results can be found in Section 5.1.3. The performance of the algorithm, along with a discussion of the relevant MSigDB terms identified within the clusters is outlined in section 5.4.

5.1.2. *Spectral Clustering*

Figure 1(b) shows the clustering results obtained when the weighted cut algorithm was applied to the directed version of the context-specific GRN. The desired number of clusters was set to 13, based on the number of clusters obtained from the MCL study, to allow for comparisons between the two algorithms. The out-degree was chosen to normalize the cut as interaction is known to be a key aspect of biological networks. Further we were interested in studying if the incorporation of direction in normalizing the cut, would provide better results compared with clustering on undirected graphs.

As seen in Figure 1(b), eight out of thirteen clusters were found to be significantly enriched with the different tumor types. The biological significance of these results is elaborated upon in Section 5.1.3.

The normalized cut algorithm (symmetric spectral clustering) was found to produce results similar to the MCL cluster assignments. Comparing the results from the symmetric clustering algorithms with asymmetric spectral clustering, we note that the clusters produced by the asymmetric variant are more balanced.

It is interesting to note that the tumor type associations derived from these cluster assignments (Figures 1(a) and (b)) corroborate the evidence obtained by Sen et al.[19] We augment the authors' findings with a further refinement of the clusters and possible associations amongst them.

5.1.3. *Biological Validation*

Following graph clustering, the obtained results were analyzed for tumor type and gene set enrichment using the methods described in Section 4.3. Significant terms were considered based on an adjusted p-value cut-off of 0.05 following which the terms were filtered based on a minimum enrichment ratio criterion of 0.1. Terms were then grouped on the basis of the source of the annotation. Pathways, Gene Ontology associations and

Table 1: Biological Significance of Clustering Results of the TN Dataset. Tumor types in bold indicate tumor types enriched with an *adjusted* p-value < 0.05. MSigDB Terms in bold indicate terms unique to the context cluster (when compared with other context clusters obtained using the same method). Acronyms in brackets indicate source of annotation (G:GenMAPP, K:KEGG, B:BioCarta, T:TRANSFAC, GO BP:Gene Ontology Biological Process, SA:SigmaAldrich, ST:Signaling Transduction KE)

Context Cluster & Tumor Type	*Relevant MSigDB Terms*
Markov Clustering	
CC2[**Pancreas**]	**Apoptosis, Purine & Pyrimidine Metabolism[G]** **Gycolysis & Gluconeogenesis*[G], Ribosome[K]** **FAS Signaling Pathway*[B]** **AP2 Family TF Binding Site[T], MTA3, RAR-RXR, MTOR Pathways[B]** Oxidative phosphorylation[G], Proteasome[K] Positive Regulation of Signal Transduction[GO BP]
CC6[Breast, Lung, Ovarian]	**Classic Pathway[B], Comp Pathway[B]** **Complement Activation Classical[G]**
CC8[Cartilage, Chondrosarcoma]	**Methane Metabolism[G], Cholera Infection[K]** **Stilbene Coumarine and Lignin Biosynthesis[G]**
CC10[Salivary, Ovarian]	**EI2F Pathway[B], Translation Factors[G]**
Spectral Clustering Asymmetric	
CC2[**Pancreas**]	**SA G1 & S Phases[SA], MTOR & PTDINS Pathway[B],** Gycolysis, RAR-RXR and MTA3 Pathway[B],
CC7[**Pancreas**]	**Gycolysis & Gluconeogenesis*[G, K] FAS Signaling Pathway*[ST]** Ribosome Pathway[K], Pyrimidine Metabolism, Proteasome[G], Regulation of JAK-STAT Cascade, Nuclear Export, RNA Splicing[GO BP]
CC12[**Pancreas**]	**TNFR1, IL2 2BP Pathway[B], Starch & Sucrose Metabolism*[G]**, **IL6 Pathway[B]**, Gycolysis Pathway*[B], MTA3 Pathway[B]
CC13[**Pancreas**]	**Actin Y Pathway[B], Translation Factors[G], ECM Pathway[B]** I Kappa B Kinase NFKappa B Cascade[GO BP], Glucose Catabolic Process*[GO BP], Proteasome[G], MTA3 Pathway, Gycolysis Pathway*[B], Oxidative phosphorylation*[G]
CC11[**Pancreas, Salivary**, ACSAL]	**BAD Pathway[B]**, **Pathogenic *E.Coli* Infection EHEC & EPEC[K],** Gycolysis Pathway*[B]
CC1[**Gall Bladder**]	Ribosome Pathway[K], Proteasome Pathway[K], mRNA Metabolic Process, mRNA processing[GO BP], Mitochondrion, Nuclear Part[GO Cellular Component], Oxidative phosphorylation[G]
CC6[**Breast**, Lung]	**Classic & Comp Pathway[B],, Complement Activation Classical[G]**
CC9[Cartilage, Chondrosarcoma]	**Methane Metabolism[G], Cholera Infection[K],** **Stilbene Coumarine and Lignin Biosynthesis[G]**

transcription factors found to be relevant to the study, along with the context clusters in which they were enriched are shown in Table 1. Context clusters enriched with tumor types but not significantly associated with MSigDB terms are not listed in the table. A complete version of the results is available on our website along with our supplementary materials.

Markov Clustering: Of the thirteen context clusters produced by MCL, CC2 and CC7 were found to be enriched with pancreatic and kidney tumors respectively (using the adjusted p-value). We also found that Symmetric Spectral clustering produced similar results. Upon closer examination of the contexts forming a part of the two clustering results, we observe that more than 90 % of the cluster assignments were identical. As seen in Table 1, CC2 (MCL) was found to be enriched with the transcription factor binding site for the AP2 family of proteins, known to play a role in the repression of pancreatic cell proliferation.[6]

Asymmetric Spectral Clustering: Tumor type enrichment analysis of the cluster assignments produced by Asymmetric Spectral Clustering yielded twelve out of the thirteen clusters enriched with different tumor types. CC6 consists of members of the HLA family; HLA-DM, whose expression when combined with that of HLA-DR, is considered to influence breast tumor progression and patient outcome.[16] Enriched subsets of the genes were also found to participate in the BRG1-induced tumor arrest in breast cancer cells.[9]

Of the remaining clusters, CC 2, 7, 11, 12, 13 were all found to be enriched with pancreatic tumor. In addition, CC11 was also enriched with salivary gland tumor, and ACSAL. It is of note that pancreatic cancer and salivary gland tumor are both tumors from secretory organs which secrete certain common enzymes (eg. Amylose). As seen in Table 1, several pathways known to play a role in tumor progression were enriched in these clusters. The *pyrimidine metabolism pathway* activity seen in CC2 could indicate a subset of pancreatic cancers which would be responsive to the agent Gemcitabine, which has been shown to improve survival for patients with pancreatic cancer (about 5 – 14 percent patients respond).

All clusters enriched with any tumor type, contained several gene sets enriched with cancer modules and cancer gene neighborhoods as defined by Brentani et al.[3] Of particular interest is the finding of a great deal of activation of metabolic genes (indicated in Table 1 by an asterisk), which is consistent with the linking of tumor cell metabolism as a target in pancreatic cancer.[21] The clinical implications of the results are further elaborated in Section 5.3

5.2. *REMBRANDT Data*

In order to test the scalability of the algorithms, and applicability to other studies, we applied the graph clustering algorithms to the REMBRANDT dataset. REpository for Molecular BRAin Neoplasia DaTa (REMBRANDT)[14] is a knowledge base consisting of clinical and functional genomics data from clinical trials involving patients suffering from gliomas. Gene expression data was collected from 417 different glioma tissue samples, and analyzed using the Affymetrix HG U133 Plus 2 microarray chip.

The raw expression values were quantized on the basis of two fold changes, and then filtered to remove transcripts with no change across all samples. Following this, the cellular context mining algorithm was applied in order to extract meaningful contexts. Statistically significant contexts (p-value < 0.0005) were then used in the construction of the graph. The resulting graph consisted of 1,901 nodes and 33,820 edges.

MCL was used with the same parameters and resulted in 32 clusters of varying sizes. Spectral clustering using the normalized-cut algorithm was applied to the undirected graph and the weighted-cut algorithm was applied to the directed graph. In both cases, the number of desired clusters was set to 32. Clusters having fewer than three nodes were not considered. Subsequently, gene and sample enrichment analysis was performed on the resulting clusters.

Tumor type enrichment and gene set enrichment analysis were conducted using the methods described in Section 4.3. Significant terms were considered based on a corrected p-value cut-off of 0.05 following which the terms were filtered based on a minimum enrichment ratio criterion of 0.1.

Table 2 shows the tumor type enrichments of the clusters obtained using MCL (on the undirected graph) and the Spectral Weighted Cut algorithm (on the directed graph). The table lists out the context clusters that were significantly enriched with each tumor type. It is to be noted that the numbering of context clusters does not match across different methods. As seen in the table we note that out of the 32 context clusters that the two algorithms produced, eleven (in the case of MCL) and ten (in the case of Spectral Asymmetric) have been enriched with tumor type associations. Interestingly, the tumor type Astrocytoma was significantly associated with Oligodendroglioma in at least two clusters.

An analysis of the MSigDB terms enriched in the clusters showed several MSigDB terms including *Cell Signaling Pathways, Cell Cycle, Cell Adhesion, Apoptosis, Regulation of DNA Replication, the E2F transcription factor pathway and the ERK Pathway.* These terms are linked to cell growth and proliferation characterizing tumor behavior. The context clusters were also found to be enriched with several cancer modules.[3] Detailed results are provided on our website along with other supplementary information.

5.3. *Clinical Implications of the Results*

The data generated by these two graph clustering methods should be of interest to clinical investigators because they have potential clinical implications. To begin with pancreatic cancer, it is comforting to see

Table 2: REMBRANDT Tumor Type Enrichment (Using Context-Specific GRN with p-value < 0.0005): Tumor types in bold indicate tumor types enriched with an *adjusted* p-value < 0.05. MSigDB Terms listed consist of terms unique to the context cluster (when compared with other context clusters obtained using the same method). Please refer to Table 1 for annotation sources. Tumor types include Glioblastoma(GBM), Astrocytoma(Astro), Oligodendroglioma(Oligo) and Mixed.

Context Cluster & Tumor Type	*Relevant MSigDB Terms*
Markov Clustering	
CC5[**GBM**]	Cell Cycle, ATR- BRCA, PLK3, P27, MPR, SKP2 E2F, G1 Pathways[B], DNA Polymerase[K] G1 to S Cell Cycle Reactome, DNA Replication[G], Pyrimidine Metabolism[K, G]
CC31[**GBM**]	HIF Pathway[B]
CC6[Astro]	IL 12, TC Apoptosis Pathway[B], Breast Cancer Estrogen Signaling[GE], Peptide GPCRS[G], Toll Like, B Cell & T Cell Receptor Signaling Pathways[K], N Glycan Degradation[G], Hematopoitic Cell Lineage, Cytokine Cytokine Receptor Interaction[K], FC Epsilon RI Signaling Pathway, Natural Killer Cell Mediated Cytotoxicity[K], JAK STAT Signaling Pathway, Arachidonic Acid Metabolism[K], Leukocyte Transendothelial Migration, N Glycan Degradation[K]
CC7[Astro]	PIP3 Signaling in B Lymphocytes[SIG], Inflam Pathway[B], IL13 Pathway[ST], PML, AS B Cell, BB Cell, IL5, SODD Pathway[B], Glycosaminoglycan Degradation[B, K] B Cell Receptor Complexes[SA], Eosinophils Pathway[B], BCR Pathway[B, SIG], B Cell Antigen Receptor[ST], Interleukin 13 Pathway[ST], Alzheimer's Disease[K],
CC12[**Astro, Oligo**]	Keratan Sulfate Biosynthesis[K]
CC14[**Astro, Oligo**]	RECK Pathway[B], ERK Pathway[B]
CC27[**Oligo**, Mixed]	IL2 2BP Pathway[B], IL10 Pathway[B]
Spectral Clustering Asymmetric	
CC11[GBM]	PML, Eosinophils, AS B Cell, IL5 Pathways[B], Prostaglandin Synthesis Regulation Pathway[G]
CC13[GBM]	ATR BRCA, PLK3, P27, G1 Pathways[B], DNA Polymerase[K], G1 to Cell Cycle Reactome[G], Pyrimidine Metabolism[G, K], DNA Replication Reactome[G], P53 Signaling Pathway[K],
CC19[Astro]	IL12, CSK, T Cytotoxic, D4GDI, NKT, CTL Pathway[B], Eicosanoid Synthesis[G], Monocyte, AMI, TC Apoptosis, Lymphocyte, CBL, T Helper & Neutrophil Pathways[B], Breast Cancer Estrogen Signaling Pathway[GE], B Cell Antigen Receptor[ST], Peptide GPCRS[G], N Glycan Degradation[G, K], GPCRDB Class B Secretin Like[G], Hematopoitic Cell Lineage[K], Cytokine Cytokine Receptor Interaction[K], T Cell, B Cell & Toll Like Receptor Signaling Pathway[K] FC Epsilon RI Signaling Pathway[K],JAK STAT Signaling Pathway[K], Natural Killer Cell Mediated Cytotoxicity[K], Arachidonic Acid Metabolism[K] Glycan Structures Degradation[K], Colorectal Cancer[K], Apoptosis[K] Leukocyte Transendothelial Migration[K], Regulation of Actin Cytoskeleton[K]
CC28[**Oligo, Astro**]	Regulation Cascade of Cyclin Expression[SA]
CC29[**Oligo**]	FAS Signaling Pathway[ST]

that the *pyrimidine pathway* appears in the results of both methods of clustering. This appearance corresponds with the pathway being a target in the disease and indeed the pathway is the target for the only drug oncologists have with some clinical activity against pancreatic cancer (it very modestly improves survival).[4] Again for pancreatic cancer the clustering data implies several metabolic pathways such as *glycolysis, gluconeogenesis, fatty acid synthase.* Since we have been so bereft of targets to go after in pancreatic cancer the present data gives us some confidence that targeting metabolic pathways in pancreatic cancer (with drugs such as phenformin) could be a very productive way to attack the disease.[27] From the Target Now clustering analyses, other possible leads for the clinic include methods to selectively go after tumor metabolism for salivary and gallbladder cancers as well.

From the clustering analysis of the REMBRANDT data, the clinical implications appear to be more limited. Concentrating on glioblastoma multiform (GBM), the worst type of brain cancer where advances are greatly needed, a possible target that appears worthy of pursuit is polo-like kinase -3. This is an important finding given the fact that *polo-like kinase inhibitors* are only now being brought into the clinic. Because of the findings in the current study, we can now include patients with GBM in the phase I study with the new PLK inhibitor NMS-1286937H.

5.4. *Performance Comparison*

In order to evaluate the clustering results obtained and compare the algorithms, we used the performance metrics – coverage and performance, as described by Brandes et al.[2]

Coverage: The coverage of a clustering C is the fraction of intra-cluster edges ($m(C)$) within the complete set of edges (m), i.e

$$\text{coverage}(C) = \frac{m(C)}{m} = \frac{m(C)}{m(C) + \overline{m(C)}} \tag{4}$$

We choose this metric as it measures the wellness of a cut in a graph by taking the edges within the cluster(s) of a graph as a fraction of all the edges. Thus, the smaller a cut, the better the coverage it would have. Both a graph with no clusters at all and a graph with several disconnected components would have a coverage of 1 due to the absence of inter-cluster edges. Sparsity of the graph would not influence the coverage as long as the intra-connectivity is much higher than the inter-connectivity. Thus we anticipate that sub-graphs created by a minimum cut would have optimal coverage.

Performance: The performance of a clustering C counts the number of "correctly interpreted" pairs of nodes in a graph. More precisely, it is the fraction of intra-cluster edges together with non-adjacent pairs of nodes in different clusters, within the set of all pairs of nodes, i.e.

$$\text{performance}(C) = \frac{m(C) + \sum_{v,w \notin E, v \in C_i, w \in C_j, i \neq j} 1}{\frac{1}{2} n(n-1)} \tag{5}$$

We choose this measure as a means to assess the connectivity within the clusters of the graph. The fewer non-edges (pairs of nodes within the same cluster but lacking an edge between them) there are within a graph, the higher its performance would be. Further, a graph containing several singleton nodes, as well as a fully connected graph with a single giant cluster, would both have a performance of 1, as the number of non-edges would be zero in both cases. The goal is to maximize connectivity within a cluster for better performance and by maximizing intra-connectivity (approaching the number of possible edges of a graph), one can minimize the inter-connectivity. Performance will not do well in sparsely connected large graphs and clusters even though there may be substantially fewer edges between clusters.

Equations 4 and 5 are specific to undirected graphs. In the case of directed graphs, the maximum number of edges possible is twice as many as the edges possible in undirected graphs and the equations are correspondingly modified.

In our first study, we compare three spectral clustering variants – symmetric spectral clustering with two variants of asymmetric spectral clustering, using different balancing parameters (the average cut and the out-degree cut). The average of performance and coverage is used as a measure of the wellness of the clusters, and is plotted against the number of clusters produced, shown in Figure 2. Spectral clustering performed well both on undirected graphs and directed graphs. We note that the asymmetric algorithms peaked at a higher number of clusters than the symmetric algorithm. This implies that the normalized cut algorithm left intact large, well connected clusters until a certain threshold was reached. We also note that using the average cut exhibits less fluctuation in performance across different cluster sizes than using the out-degree of the nodes, explained by the fact that the average cut uses the number of nodes as the balancing parameter. However, if in fact a GRN follows a scale-free topology then the average cut may not prove to be the most useful in identifying biologically significant clusters because it does not take into account the interactions within a cluster.

In our second study, we compare spectral clustering (symmetric and asymmetric) with Markov clustering. As seen in Table 3, in terms of coverage, spectral clustering performed well over both directed and undirected graphs. In terms of performance, we find that the asymmetric case shows a lower performance value than the

Table 3: Performance Comparison of Markov and Spectral Clustering

Metric	*Dataset*	*Spectral Asym.*	*Spectral Sym.*	*MCL*	*Dataset*	*Spectral Asym.*	*Spectral Sym.*	*MCL*
Coverage	TN	0.9533	0.9693	0.9366	REM	0.7144	0.9680	0.9386
Performance	TN	0.6038	0.7930	0.7696	REM	0.8804	0.9271	0.8914

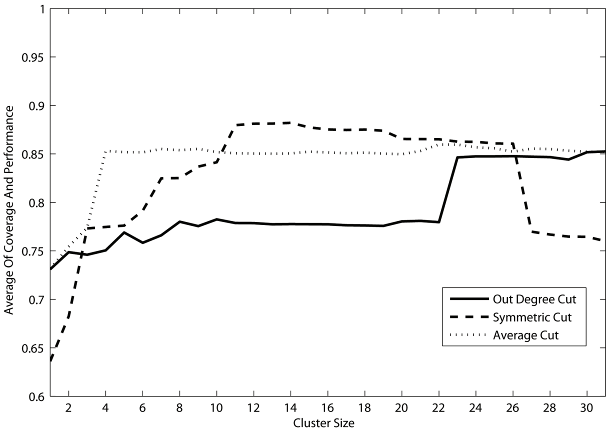

Fig. 2: Performance and Coverage Average of Spectral Clustering

other two. We also note that both MCL and symmetric spectral clustering performed well on the much larger REMBRANDT dataset, exhibiting good scalability. Comparing these results with the enrichment analyses (Tables 1 and 2, we conclude that well-balanced clusters need not necessarily correspond with biological meaningful clusters. Further we also observe that incorporating directionality did not correspond with a significant impact on the clustering, in terms of both biological significance and performance metrics.

6. Conclusion

The main contributions of our paper have been in a novel application of graph clustering, namely to identify clusters in context-specific GRNs. We have used Markov clustering and spectral clustering to identify context clusters in a two gene expression studies. The methods were compared to assess their ability to produce disjoint balanced clusters and scale to large graphs. Functional annotation of the genes and sample association studies show graph clustering to be promising in this area.

Future work includes studying the cluster enrichments obtained at increasing levels of cluster granularity, as well as the incorporation of prior biological knowledge into the clustering framework.

Acknowledgments

The authors wish to acknowledge Ina Sen, Michael Verdicchio and Sara Nasser for their help in the construction of the graphs and preparation of the manuscript. AR was supported by NIH R21 LM009706, SK was partially supported by NIH LM009706, CAA 0243-08 and P01 CA109552, and DVH by NIH P01 CA109552.

References

1. Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.
2. U. Brandes, M. Gaertler, and D. Wagner. Experiments On Graph Clustering Algorithms. *Lecture Notes in Computer Science*, pages 568–579, 2003.
3. H. Brentani, O. Caballero, A. Camargo, A. da Silva, W. da Silva, E. Neto, M. Grivet, A. Gruber, P. Guimaraes, W. Hide, et al. The Generation and Utilization of a Cancer-Oriented Representation of the Human Transcriptome by Using Expressed Sequence Tags. *Proceedings of the National Academy of Sciences*, 100(23):13418–13423, 2003.

4. H. Burris 3rd, M. Moore, J. Andersen, M. Green, M. Rothenberg, M. Modiano, M. Cripps, R. Portenoy, A. Storniolo, P. Tarassoff, et al. Improvements in Survival and Clinical Benefit with Gemcitabine As First-Line Therapy for Patients with Advanced Pancreas Cancer: A Randomized Trial. *Journal of Clinical Oncology*, 15(6):2403, 1997.
5. R. Dubes and A. Jain. Algorithms for Clustering Data. *Prentice Hall*, 355:356, 1988.
6. V. Fauquette, S. Aubert, S. Groux-Degroote, B. Hemon, N. Porchet, I. Van Seuningen, and P. Pigny. Transcription Factor AP-2 {Alpha} Represses Both the Mucin MUC4 Expression and Pancreatic Cancer Cell Proliferation. *Carcinogenesis*, 28(11):2305, 2007.
7. T. Freeman, L. Goldovsky, M. Brosch, S. van Dongen, P. Mazière, R. Grocock, S. Freilich, J. Thornton, and A. Enright. Construction, Visualisation, and Clustering of Transcription Networks From Microarray Expression Data. *PLoS Comput Biol*, 3(10):2032–2042, 2007.
8. W. Hahn and R. Weinberg. Modelling the Molecular Circuitry of Cancer. *Nat. Rev. Cancer*, 2(5):331–341, 2002.
9. K. Hendricks, F. Shanahan, and E. Lees. Role for BRG1 in Cell Cycle Control and Tumor Suppression. *Molecular and Cellular Biology*, 24(1):362–376, 2004.
10. D. Higham, G. Kalna, and M. Kibble. Spectral Clustering and Its Use in Bioinformatics. *Journal of computational and applied mathematics*, 204(1):25–37, 2007.
11. A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3), 1999.
12. S. Kim, I. Sen, and M. Bittner. Mining Molecular Contexts of Cancer Via in-Silico Conditioning. In *Computational Systems Bioinformatics: Proceedings of the CSB 2007 Conference*. Imperial College Press, 2007.
13. Y. Kluger, R. Basri, J. Chang, and M. Gerstein. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions, 2003.
14. S. Madhavan, J. Zenklusen, Y. Kotliarov, H. Sahni, H. Fine, and K. Buetow. Rembrandt: Helping Personalized Medicine Become a Reality Through Integrative Translational Research. *Mole. Cancer Res.*, 7(2):157, 2009.
15. M. Meila and W. Pentney. Clustering by Weighted Cuts in Directed Graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, 2007.
16. S. Oldford, J. Robb, D. Codner, V. Gadag, P. Watson, and S. Drover. Tumor Cell Expression of HLA-DM Associates with a Th1 Profile and Predicts Improved Survival in Breast Carcinoma Patients. *International immunology*, 18(11):1591, 2006.
17. B. Samuel Lattimore, S. van Dongen, and M. Crabbe. GeneMCL in Microarray Analysis. *Computational Biology and Chemistry*, 29(5):354–359, 2005.
18. S. Schaeffer. Graph Clustering. *Computer Science Review*, 1(1):27–64, 2007.
19. I. Sen, M. Verdicchio, S. Jung, R. Trevino, M. Bittner, and S. Kim. Context-Specific Gene Regulations in Cancer. In *Proceedings of the Pacific Symposium on Biocomputing*, 2009.
20. J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 888–905, 2000.
21. J. Spratlin, N. Serkova, and S. Eckhardt. Clinical Applications of Metabolomics in Oncology: A Review. *Clinical Cancer Research*, 15(2):431, 2009.
22. A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
23. D. Tritchler, S. Fallah, and J. Beyene. A Spectral Clustering Method for Microarray Data. *Computational Statistics and Data Analysis*, 49(1):63–76, 2005.
24. S. van Dongen. Graph Clustering by Flow Simulation. *University of Utrecht*, 2000.
25. S. van Dongen. Technical Report INS-R0010: A Cluster Algorithm for Graphs. *National Research Institute for Mathematics and Computer Science*, 2000.
26. D. Von Hoff, R. Penny, S. Shack, E. Campbell, D. Taverna, M. Borad, D. Love, J. Trent, and M. Bittner. Frequency of Potential Therapeutic Targets Identified by Immunohistochemistry (IHC) and DNA Microarray (DMA) in Tumors From Patients Who Have Progressed On Multiple Therapeutic Agents. *Journal of Clinical Oncology*, 24(18_suppl):3071, 2006.
27. D. Wise, R. DeBerardinis, A. Mancuso, N. Sayed, X. Zhang, H. Pfeiffer, I. Nissim, E. Daikhin, M. Yudkoff, S. McMahon, et al. Myc Regulates a Transcriptional Program That Stimulates Mitochondrial Glutaminolysis and Leads to Glutamine Addiction. *Proceedings of the National Academy of Sciences*, 105(48):18782, 2008.